UniClever

Text Analysis Software

Ian Johnson, Julien Maudet

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Unilever

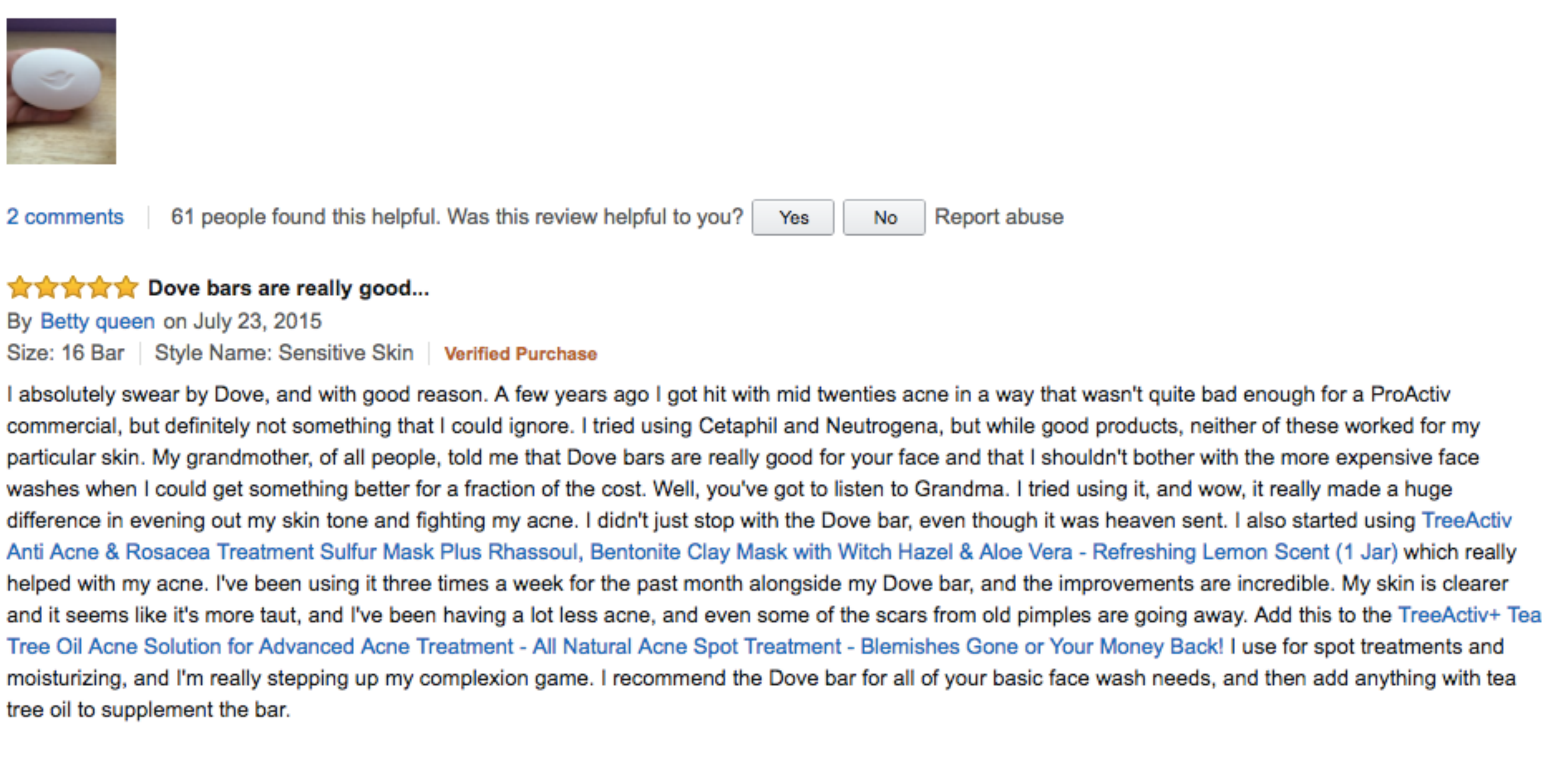
Context

- Unilever looking for useful analysis of their survey and reviews data for beauty products

- End product to be an interface for analysts to gather quick summarization / insight into large numbers of reviews or survey answers

Data

- 198,502 Amazon Reviews



2 comments | 61 people found this helpful. Was this review helpful to you? Report abuse

★★★★★ **Dove bars are really good...**

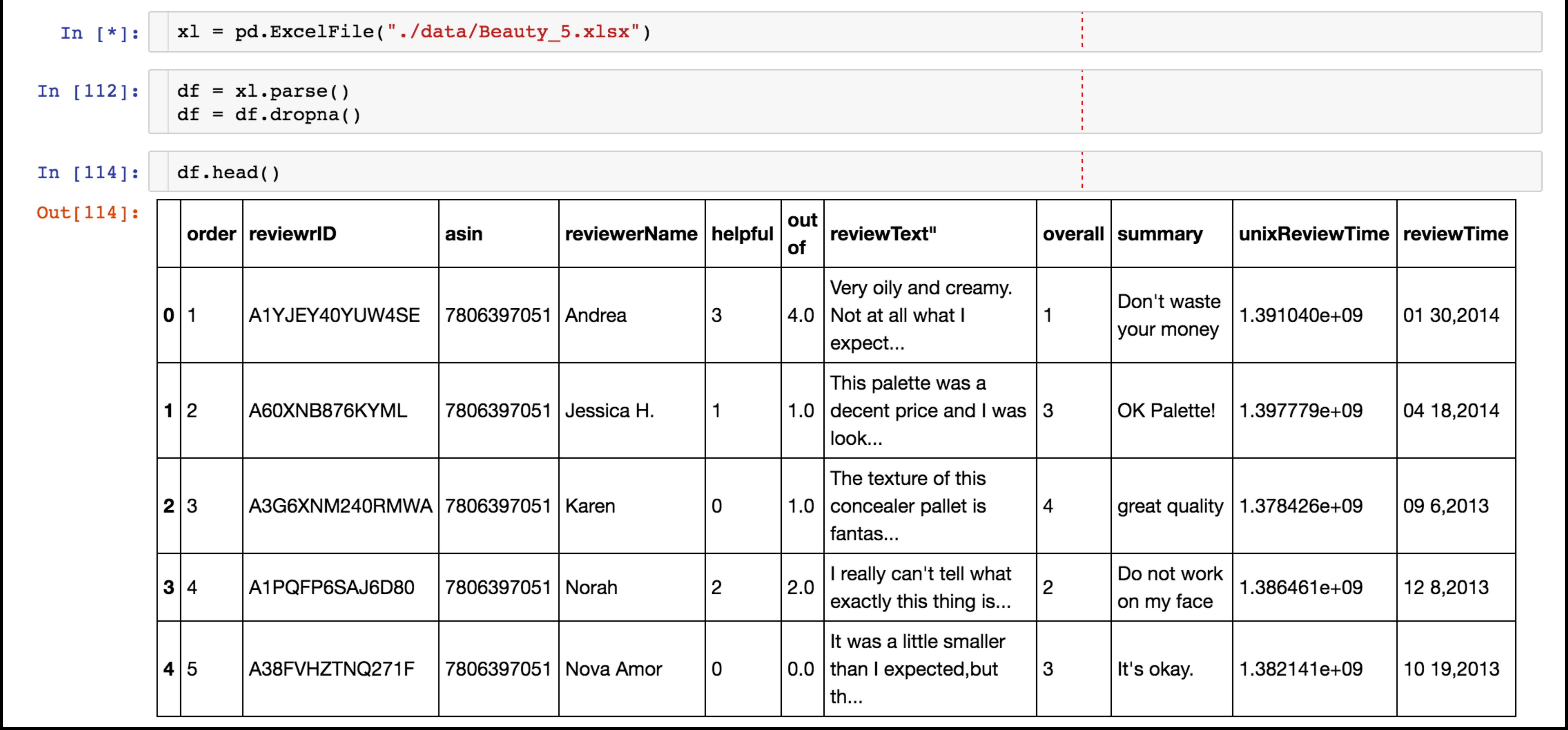
By Betty queen on July 23, 2015

Size: 16 Bar | Style Name: Sensitive Skin | **Verified Purchase**

I absolutely swear by Dove, and with good reason. A few years ago I got hit with mid twenties acne in a way that wasn't quite bad enough for a ProActiv commercial, but definitely not something that I could ignore. I tried using Cetaphil and Neutrogena, but while good products, neither of these worked for my particular skin. My grandmother, of all people, told me that Dove bars are really good for your face and that I shouldn't bother with the more expensive face washes when I could get something better for a fraction of the cost. Well, you've got to listen to Grandma. I tried using it, and wow, it really made a huge difference in evening out my skin tone and fighting my acne. I didn't just stop with the Dove bar, even though it was heaven sent. I also started using TreeActiv Anti Acne & Rosacea Treatment Sulfur Mask Plus Rhassoul, Bentonite Clay Mask with Witch Hazel & Aloe Vera - Refreshing Lemon Scent (1 Jar) which really helped with my acne. I've been using it three times a week for the past month alongside my Dove bar, and the improvements are incredible. My skin is clearer and it seems like it's more taut, and I've been having a lot less acne, and even some of the scars from old pimples are going away. Add this to the TreeActiv+ Tea Tree Oil Acne Solution for Advanced Acne Treatment - All Natural Acne Spot Treatment - Blemishes Gone or Your Money Back! I use for spot treatments and moisturizing, and I'm really stepping up my complexion game. I recommend the Dove bar for all of your basic face wash needs, and then add anything with tea tree oil to supplement the bar.

Data

- 198,502 Amazon Reviews



In [*]:

```
xl = pd.ExcelFile("./data/Beauty_5.xlsx")
```

In [112]:

```
df = xl.parse()
df = df.dropna()
```

In [114]:

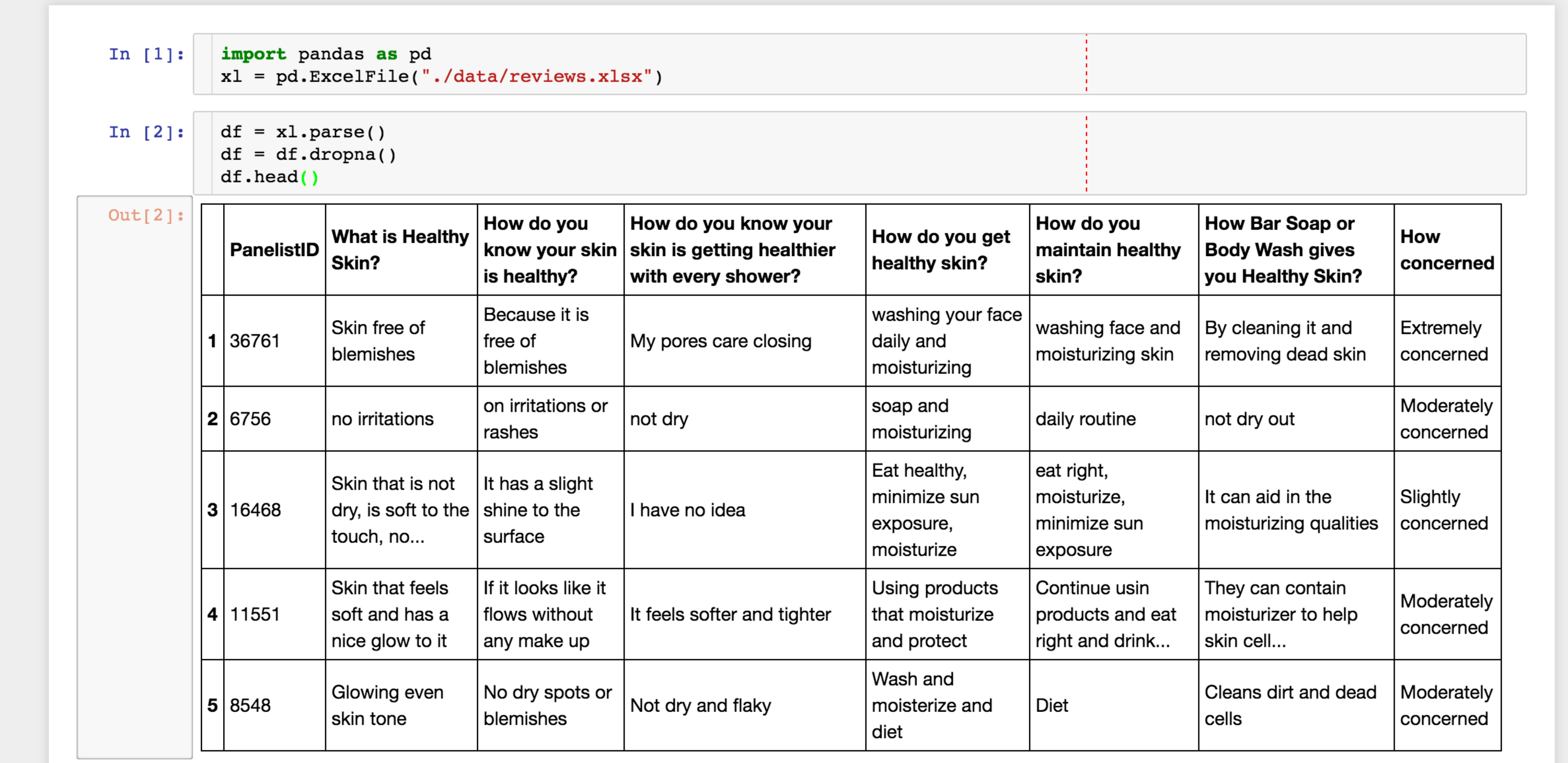
```
df.head()
```

Out[114]:

	order	reviewrID	asin	reviewerName	helpful	out of	reviewText"	overall	summary	unixReviewTime	reviewTime
0	1	A1YJEY40YUW4SE	7806397051	Andrea	3	4.0	Very oily and creamy. Not at all what I expect...	1	Don't waste your money	1.391040e+09	01 30,2014
1	2	A60XNB876KYML	7806397051	Jessica H.	1	1.0	This palette was a decent price and I was look...	3	OK Palette!	1.397779e+09	04 18,2014
2	3	A3G6XNM240RMWA	7806397051	Karen	0	1.0	The texture of this concealer pallet is fantas...	4	great quality	1.378426e+09	09 6,2013
3	4	A1PQFP6SAJ6D80	7806397051	Norah	2	2.0	I really can't tell what exactly this thing is...	2	Do not work on my face	1.386461e+09	12 8,2013
4	5	A38FVHZTNQ271F	7806397051	Nova Amor	0	0.0	It was a little smaller than I expected,but th...	3	It's okay.	1.382141e+09	10 19,2013

Data

- 198,468 Survey Answers



```
In [1]: import pandas as pd
        xl = pd.ExcelFile("./data/reviews.xlsx")
```

```
In [2]: df = xl.parse()
        df = df.dropna()
        df.head()
```

Out[2]:

	PanelistID	What is Healthy Skin?	How do you know your skin is healthy?	How do you know your skin is getting healthier with every shower?	How do you get healthy skin?	How do you maintain healthy skin?	How Bar Soap or Body Wash gives you Healthy Skin?	How concerned
1	36761	Skin free of blemishes	Because it is free of blemishes	My pores care closing	washing your face daily and moisturizing	washing face and moisturizing skin	By cleaning it and removing dead skin	Extremely concerned
2	6756	no irritations	on irritations or rashes	not dry	soap and moisturizing	daily routine	not dry out	Moderately concerned
3	16468	Skin that is not dry, is soft to the touch, no...	It has a slight shine to the surface	I have no idea	Eat healthy, minimize sun exposure, moisturize	eat right, moisturize, minimize sun exposure	It can aid in the moisturizing qualities	Slightly concerned
4	11551	Skin that feels soft and has a nice glow to it	If it looks like it flows without any make up	It feels softer and tighter	Using products that moisturize and protect	Continue usin products and eat right and drink...	They can contain moisturizer to help skin cell...	Moderately concerned
5	8548	Glowing even skin tone	No dry spots or blemishes	Not dry and flaky	Wash and moisterize and diet	Diet	Cleans dirt and dead cells	Moderately concerned

Summarization

Overview

Split sentences from collection of documents —>

—> Chunk sentences ? —>

—> (Ignore Misspellings) ? —>

—> (Split Sibling Sentences) ? —>

—> Lemmatization —>

—> Remove Stopword bigrams —>

—> [Word Count | Tfidf Vectorization] (n-gram range) —>

—> (Normalization) —>

—> SVD (k) —>

—> Semantic Volume Maximization (Yogotama et al.) —>

—> Return Summary Sentences

() — Locally implemented / untested

? — Optional step

Summarization

SVD for Text Summarization

Input: Tfidf-weighted or word count vectors for each sentence (after preprocessing)

1. Choose k, the dimension of the space onto which to map the vectorized sentences (e.g. the number of "topics")

2. Vectorized sentences become jointly indexed by the singular vectors of this lower dimensional space

   (e.g. each sentence vector becomes represented by a combination of the k singular vectors)

3. Magnitudes of the values of the mapped sentences represent their correspondence with each of these singular vectors

—> This result will be leveraged by the Semantic Volume Maximization algorithm

1. Steinberger, Josef, and Karel Ježek. 2004. Text summarization and singular value decomposition. Iternational Conference on Advances in Information Systems. Springer Berlin Heidelberg.

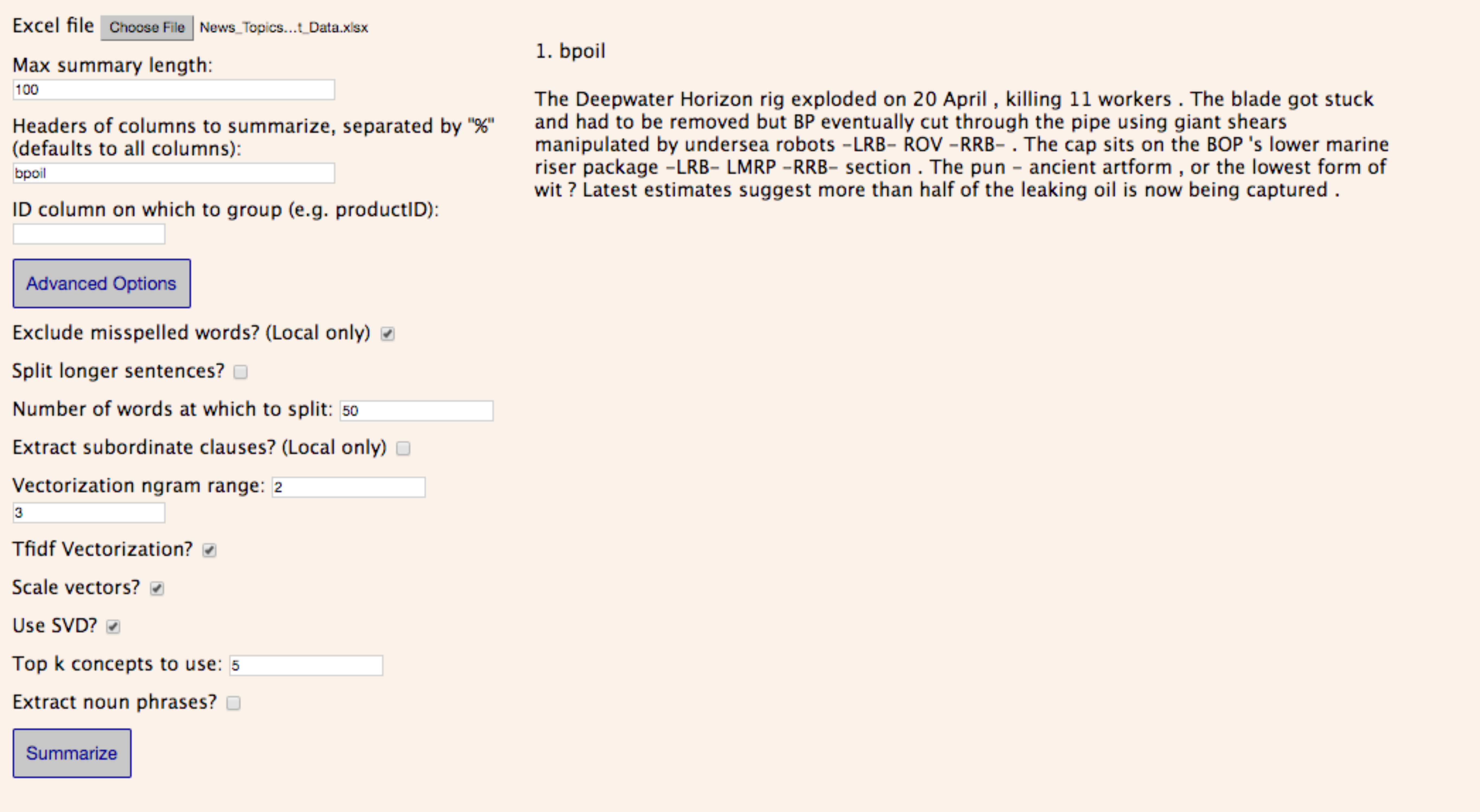
Summarization

Semantic Volume Maximization

- SVD creates a representation of the sentences in terms of their correspondences to the top “topics” of the corpus (e.g. condenses the semantic space)
- Now we seek to choose the sentences whose vectors will together maximize the volume of this condensed space:

- Input: Summary length w (# of words or sentences)
    - Initialize basis vector set B, result sentence set S
    - Calculate mean vector c across all sentences
    - Choose first basis vector b1 as vector most distant from c
    - Choose second basis vector b2 as vector most distant from b1 —> add it to B
    - While total length of sentences in S < w:
        - Calculate sentence vector r most distant from subspace spanned by basis vectors in B
        - Add r to B
        - Add sentence corresponding to r to S

Summarization



Excel file Choose File News_Topics...t_Data.xlsx

Max summary length: 100

Headers of columns to summarize, separated by "%" (defaults to all columns): bpoil

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? □

Number of words at which to split: 50

Extract subordinate clauses? (Local only) □

Vectorization ngram range: 2 3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

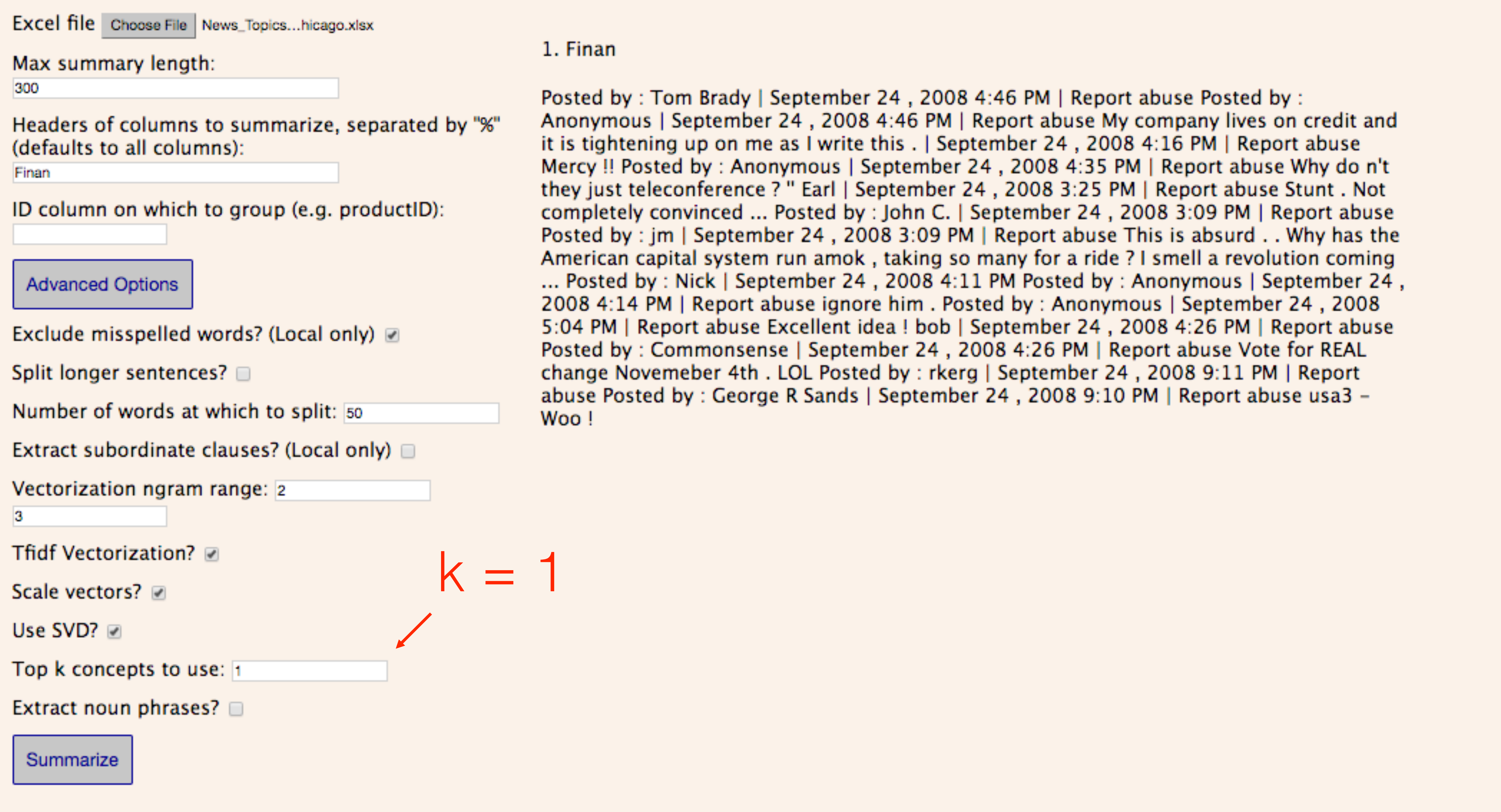
Top k concepts to use: 5

Extract noun phrases? □

1. bpoil

The Deepwater Horizon rig exploded on 20 April , killing 11 workers . The blade got stuck and had to be removed but BP eventually cut through the pipe using giant shears manipulated by undersea robots -LRB- ROV -RRB- . The cap sits on the BOP 's lower marine riser package -LRB- LMRP -RRB- section . The pun - ancient artform , or the lowest form of wit ? Latest estimates suggest more than half of the leaking oil is now being captured .

Summarization



Excel file News_Topics...hicago.xlsx

Max summary length: 300

Headers of columns to summarize, separated by "%" (defaults to all columns): Finan

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? □

Number of words at which to split: 50

Extract subordinate clauses? (Local only) □

Vectorization ngram range: 2
3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

Top k concepts to use: 1

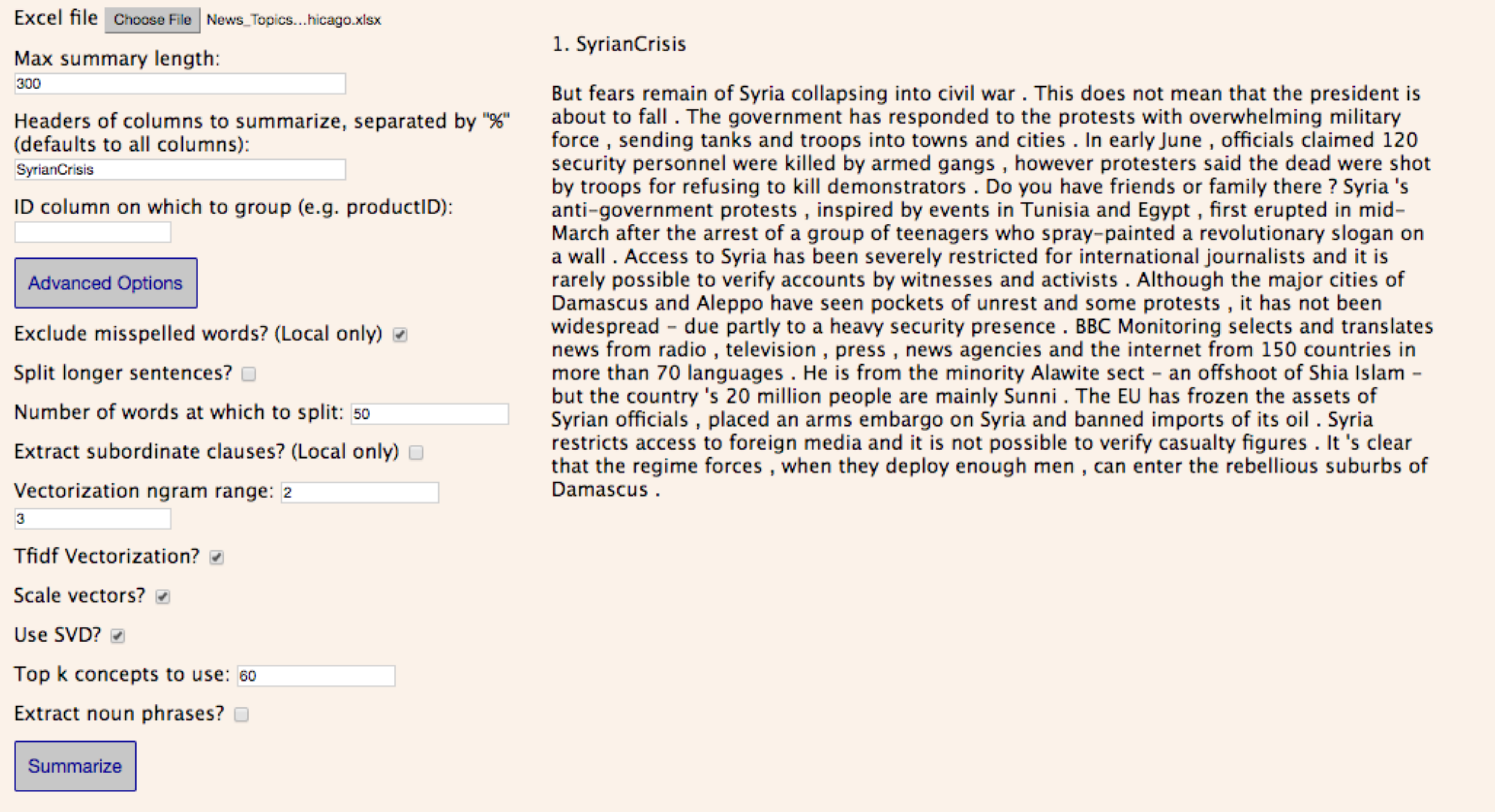
Extract noun phrases? □

k = 1

1. Finan

Posted by : Tom Brady | September 24 , 2008 4:46 PM | Report abuse Posted by : Anonymous | September 24 , 2008 4:46 PM | Report abuse My company lives on credit and it is tightening up on me as I write this . | September 24 , 2008 4:16 PM | Report abuse Mercy !! Posted by : Anonymous | September 24 , 2008 4:35 PM | Report abuse Why do n't they just teleconference ? " Earl | September 24 , 2008 3:25 PM | Report abuse Stunt . Not completely convinced ... Posted by : John C. | September 24 , 2008 3:09 PM | Report abuse Posted by : jm | September 24 , 2008 3:09 PM | Report abuse This is absurd . . Why has the American capital system run amok , taking so many for a ride ? I smell a revolution coming ... Posted by : Nick | September 24 , 2008 4:11 PM Posted by : Anonymous | September 24 , 2008 4:14 PM | Report abuse ignore him . Posted by : Anonymous | September 24 , 2008 5:04 PM | Report abuse Excellent idea ! bob | September 24 , 2008 4:26 PM | Report abuse Posted by : Commonsense | September 24 , 2008 4:26 PM | Report abuse Vote for REAL change Novemeber 4th . LOL Posted by : rkerg | September 24 , 2008 9:11 PM | Report abuse Posted by : George R Sands | September 24 , 2008 9:10 PM | Report abuse usa3 – Woo !

Summarization



Excel file News_Topics...hicago.xlsx

Max summary length: 300

Headers of columns to summarize, separated by "%" (defaults to all columns): SyrianCrisis

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? ☐

Number of words at which to split: 50

Extract subordinate clauses? (Local only) ☐

Vectorization ngram range: 2
3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

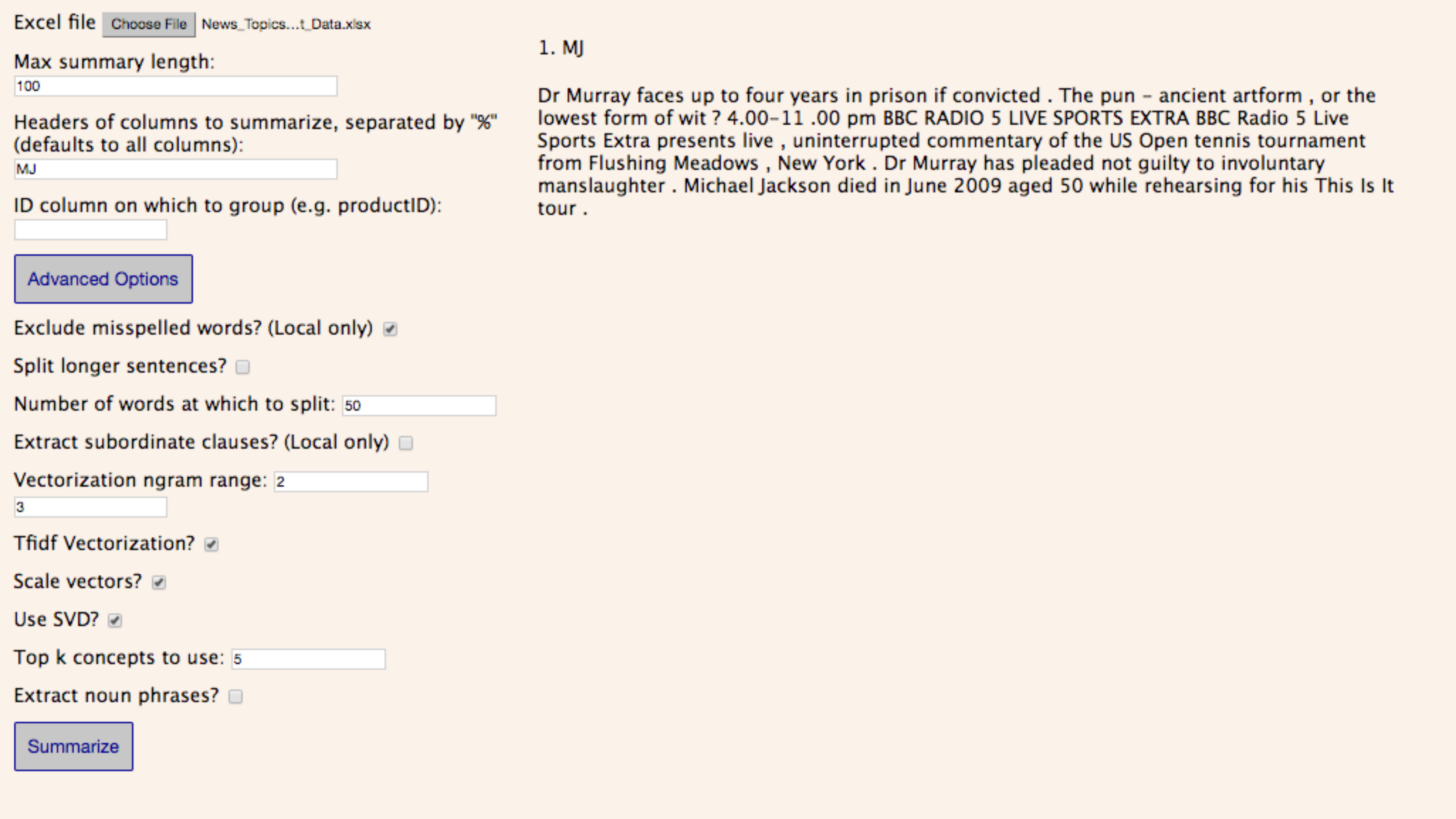
Top k concepts to use: 60

Extract noun phrases? ☐

1. SyrianCrisis

But fears remain of Syria collapsing into civil war . This does not mean that the president is about to fall . The government has responded to the protests with overwhelming military force , sending tanks and troops into towns and cities . In early June , officials claimed 120 security personnel were killed by armed gangs , however protesters said the dead were shot by troops for refusing to kill demonstrators . Do you have friends or family there ? Syria 's anti-government protests , inspired by events in Tunisia and Egypt , first erupted in mid-March after the arrest of a group of teenagers who spray-painted a revolutionary slogan on a wall . Access to Syria has been severely restricted for international journalists and it is rarely possible to verify accounts by witnesses and activists . Although the major cities of Damascus and Aleppo have seen pockets of unrest and some protests , it has not been widespread - due partly to a heavy security presence . BBC Monitoring selects and translates news from radio , television , press , news agencies and the internet from 150 countries in more than 70 languages . He is from the minority Alawite sect - an offshoot of Shia Islam - but the country 's 20 million people are mainly Sunni . The EU has frozen the assets of Syrian officials , placed an arms embargo on Syria and banned imports of its oil . Syria restricts access to foreign media and it is not possible to verify casualty figures . It 's clear that the regime forces , when they deploy enough men , can enter the rebellious suburbs of Damascus .

Summarization



Excel file News_Topics...t_Data.xlsx

Max summary length: 100

Headers of columns to summarize, separated by "%" (defaults to all columns): MJ

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? □

Number of words at which to split: 50

Extract subordinate clauses? (Local only) □

Vectorization ngram range: 2 3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

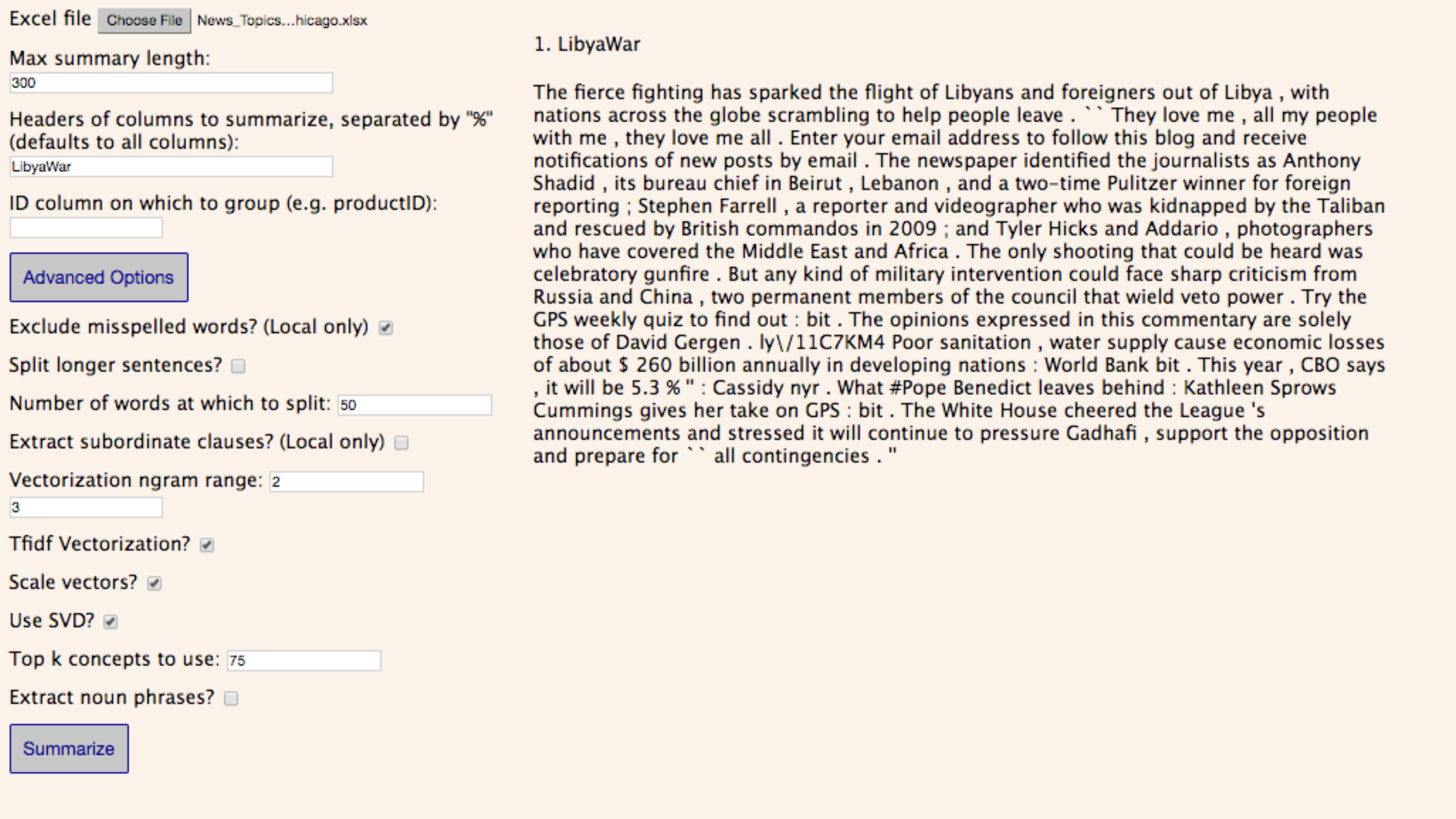
Top k concepts to use: 5

Extract noun phrases? □

1. MJ

Dr Murray faces up to four years in prison if convicted . The pun – ancient artform , or the lowest form of wit ? 4.00–11 .00 pm BBC RADIO 5 LIVE SPORTS EXTRA BBC Radio 5 Live Sports Extra presents live , uninterrupted commentary of the US Open tennis tournament from Flushing Meadows , New York . Dr Murray has pleaded not guilty to involuntary manslaughter . Michael Jackson died in June 2009 aged 50 while rehearsing for his This Is It tour .

Summarization



Excel file News_Topics...hicago.xlsx

Max summary length: 300

Headers of columns to summarize, separated by "%" (defaults to all columns): LibyaWar

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? □

Number of words at which to split: 50

Extract subordinate clauses? (Local only) □

Vectorization ngram range: 2 3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

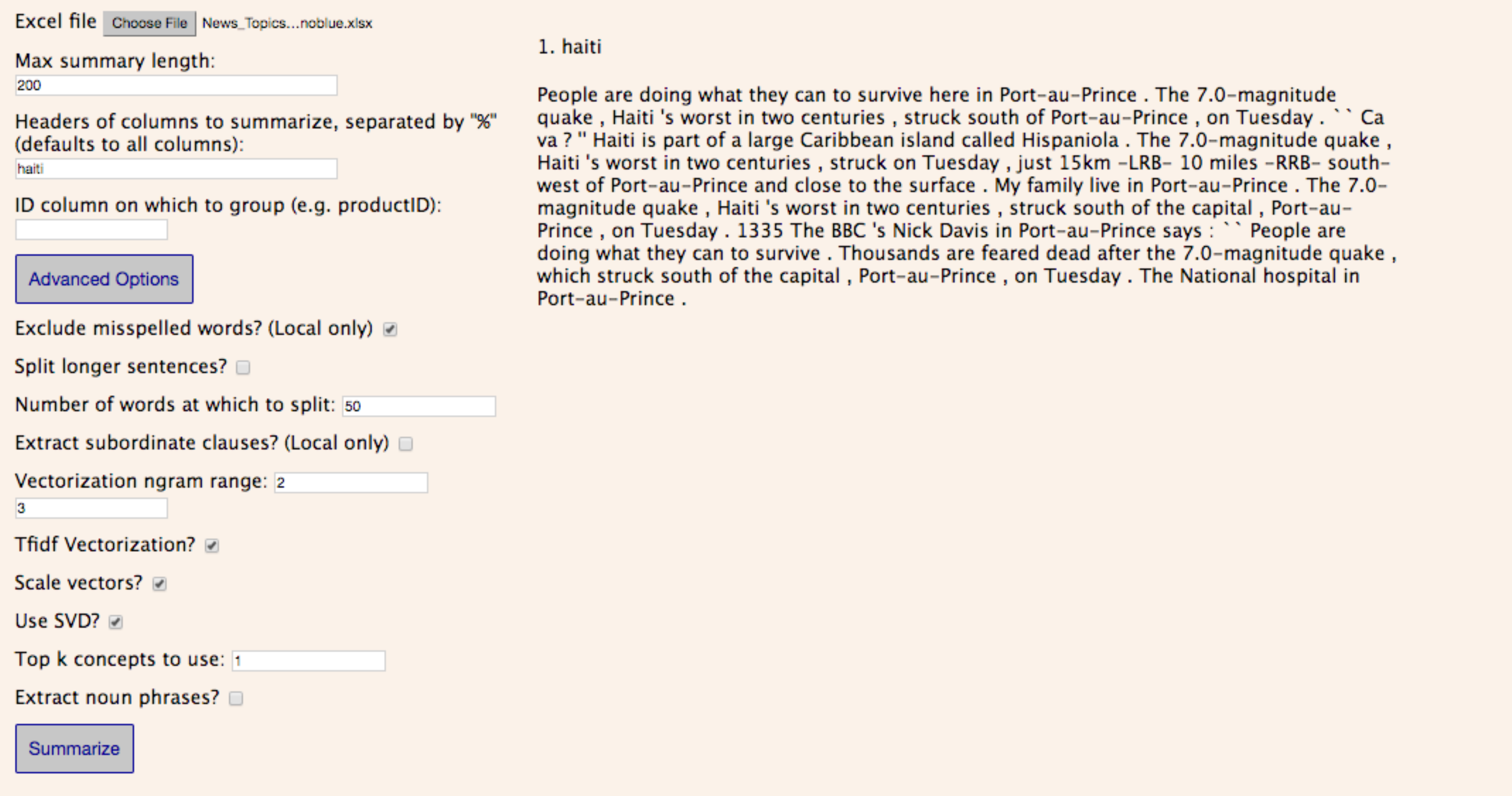
Top k concepts to use: 75

Extract noun phrases? □

1. LibyaWar

The fierce fighting has sparked the flight of Libyans and foreigners out of Libya , with nations across the globe scrambling to help people leave . `` They love me , all my people with me , they love me all . Enter your email address to follow this blog and receive notifications of new posts by email . The newspaper identified the journalists as Anthony Shadid , its bureau chief in Beirut , Lebanon , and a two-time Pulitzer winner for foreign reporting ; Stephen Farrell , a reporter and videographer who was kidnapped by the Taliban and rescued by British commandos in 2009 ; and Tyler Hicks and Addario , photographers who have covered the Middle East and Africa . The only shooting that could be heard was celebratory gunfire . But any kind of military intervention could face sharp criticism from Russia and China , two permanent members of the council that wield veto power . Try the GPS weekly quiz to find out : bit . The opinions expressed in this commentary are solely those of David Gergen . ly\/11C7KM4 Poor sanitation , water supply cause economic losses of about $ 260 billion annually in developing nations : World Bank bit . This year , CBO says , it will be 5.3 % " : Cassidy nyr . What #Pope Benedict leaves behind : Kathleen Sprows Cummings gives her take on GPS : bit . The White House cheered the League 's announcements and stressed it will continue to pressure Gadhafi , support the opposition and prepare for `` all contingencies . "

Summarization



Excel file News_Topics...noblue.xlsx

Max summary length: 200

Headers of columns to summarize, separated by "%" (defaults to all columns): haiti

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? □

Number of words at which to split: 50

Extract subordinate clauses? (Local only) □

Vectorization ngram range: 2
3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

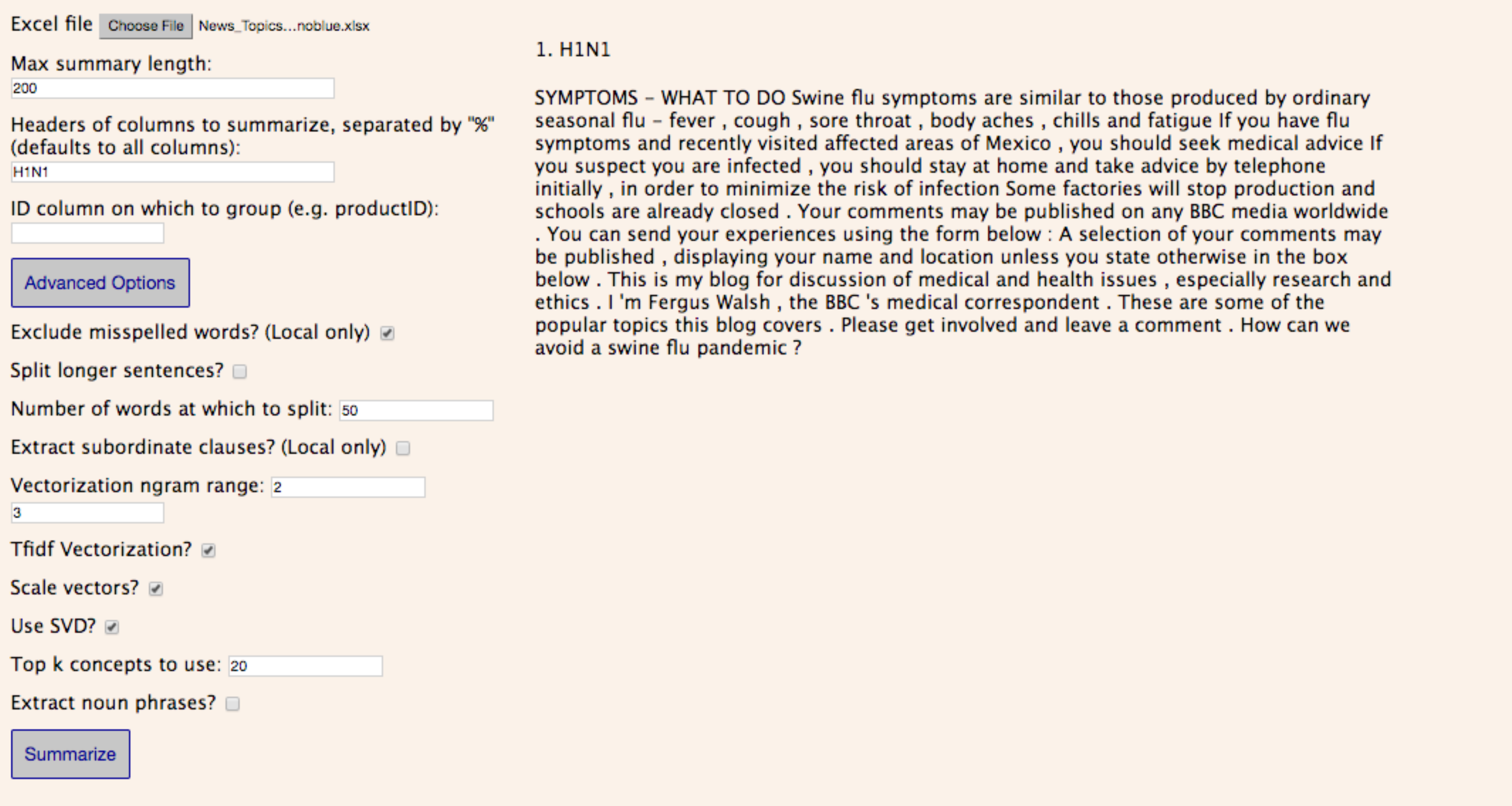
Top k concepts to use: 1

Extract noun phrases? □

1. haiti

People are doing what they can to survive here in Port-au-Prince . The 7.0-magnitude quake , Haiti 's worst in two centuries , struck south of Port-au-Prince , on Tuesday . `` Ca va ? '' Haiti is part of a large Caribbean island called Hispaniola . The 7.0-magnitude quake , Haiti 's worst in two centuries , struck on Tuesday , just 15km -LRB- 10 miles -RRB- south-west of Port-au-Prince and close to the surface . My family live in Port-au-Prince . The 7.0-magnitude quake , Haiti 's worst in two centuries , struck south of the capital , Port-au-Prince , on Tuesday . 1335 The BBC 's Nick Davis in Port-au-Prince says : `` People are doing what they can to survive . Thousands are feared dead after the 7.0-magnitude quake , which struck south of the capital , Port-au-Prince , on Tuesday . The National hospital in Port-au-Prince .

Summarization



Excel file News_Topics...noblue.xlsx

Max summary length: 200

Headers of columns to summarize, separated by "%" (defaults to all columns): H1N1

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? ☐

Number of words at which to split: 50

Extract subordinate clauses? (Local only) ☐

Vectorization ngram range: 2 3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

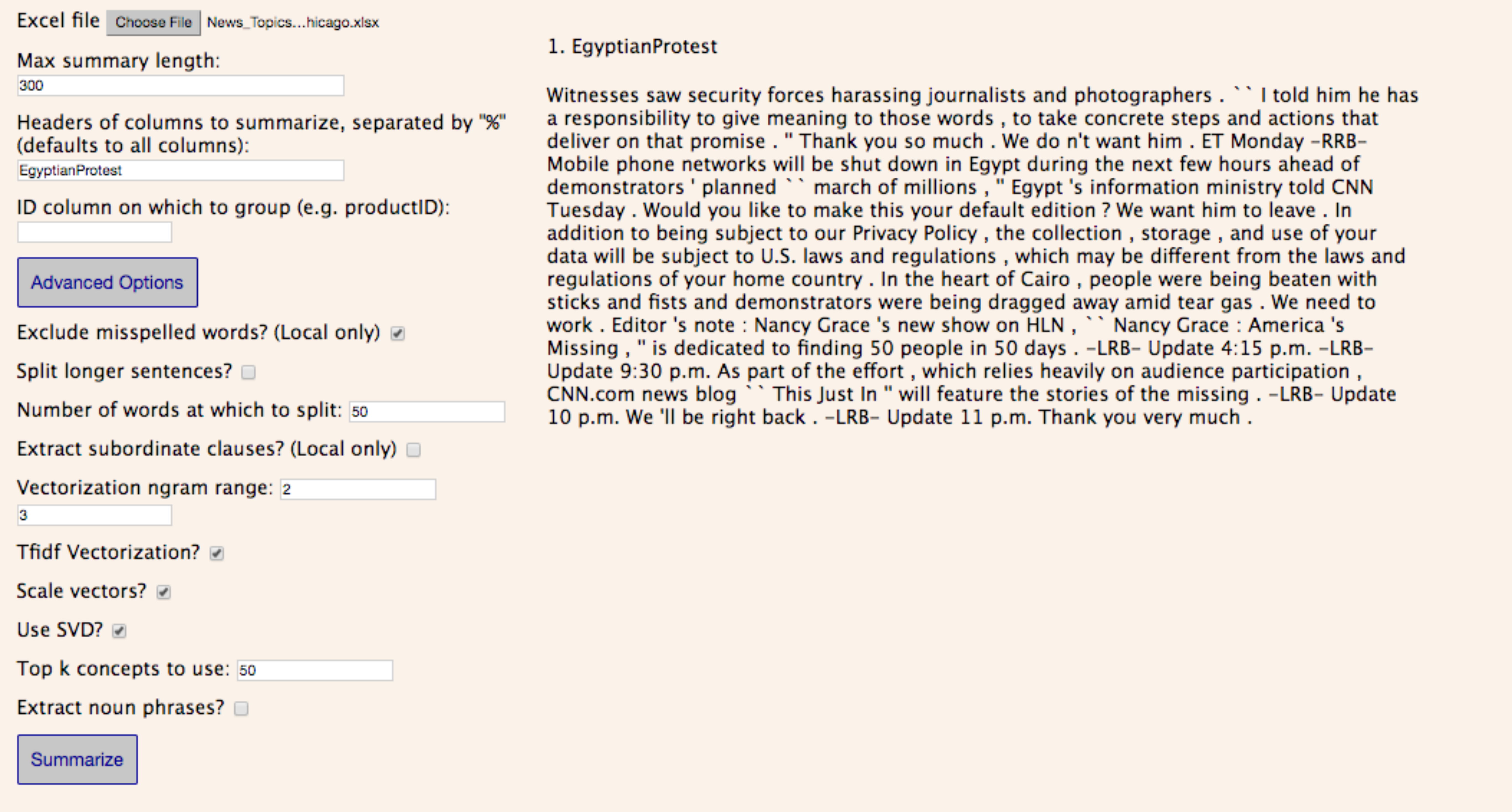
Top k concepts to use: 20

Extract noun phrases? ☐

1. H1N1

SYMPTOMS – WHAT TO DO Swine flu symptoms are similar to those produced by ordinary seasonal flu – fever , cough , sore throat , body aches , chills and fatigue If you have flu symptoms and recently visited affected areas of Mexico , you should seek medical advice If you suspect you are infected , you should stay at home and take advice by telephone initially , in order to minimize the risk of infection Some factories will stop production and schools are already closed . Your comments may be published on any BBC media worldwide . You can send your experiences using the form below : A selection of your comments may be published , displaying your name and location unless you state otherwise in the box below . This is my blog for discussion of medical and health issues , especially research and ethics . I 'm Fergus Walsh , the BBC 's medical correspondent . These are some of the popular topics this blog covers . Please get involved and leave a comment . How can we avoid a swine flu pandemic ?

Summarization



Excel file News_Topics...hicago.xlsx

Max summary length: 300

Headers of columns to summarize, separated by "%" (defaults to all columns): EgyptianProtest

ID column on which to group (e.g. productID):

Exclude misspelled words? (Local only) ☑

Split longer sentences? □

Number of words at which to split: 50

Extract subordinate clauses? (Local only) □

Vectorization ngram range: 2 3

Tfidf Vectorization? ☑

Scale vectors? ☑

Use SVD? ☑

Top k concepts to use: 50

Extract noun phrases? □

1. EgyptianProtest

Witnesses saw security forces harassing journalists and photographers . `` I told him he has a responsibility to give meaning to those words , to take concrete steps and actions that deliver on that promise . '' Thank you so much . We do n't want him . ET Monday -RRB- Mobile phone networks will be shut down in Egypt during the next few hours ahead of demonstrators ' planned `` march of millions , '' Egypt 's information ministry told CNN Tuesday . Would you like to make this your default edition ? We want him to leave . In addition to being subject to our Privacy Policy , the collection , storage , and use of your data will be subject to U.S. laws and regulations , which may be different from the laws and regulations of your home country . In the heart of Cairo , people were being beaten with sticks and fists and demonstrators were being dragged away amid tear gas . We need to work . Editor 's note : Nancy Grace 's new show on HLN , `` Nancy Grace : America 's Missing , '' is dedicated to finding 50 people in 50 days . -LRB- Update 4:15 p.m. -LRB- Update 9:30 p.m. As part of the effort , which relies heavily on audience participation , CNN.com news blog `` This Just In '' will feature the stories of the missing . -LRB- Update 10 p.m. We 'll be right back . -LRB- Update 11 p.m. Thank you very much .

Keyphrase extraction

RAKE algorithm

Source: Rose, Stuart, et al. "Automatic keyword extraction from individual documents." *Text Mining* (2010): 1-20.

Unsupervised & independent from the language at use

- ⚡ More computationally efficient than TextRank

- ⚡ Higher precision and comparable recall scores

Keyphrase extraction

RAKE algorithm

First observation

- keywords = multiple words but no punctuation or stop words (and, the…)

Input

- Document
- List of stop words and phrase delimiters
- Parameters:
  - minimum length of a word in a keyphrase
  - minimum frequency for a word in the text
  - maximum number of words per keyphrase

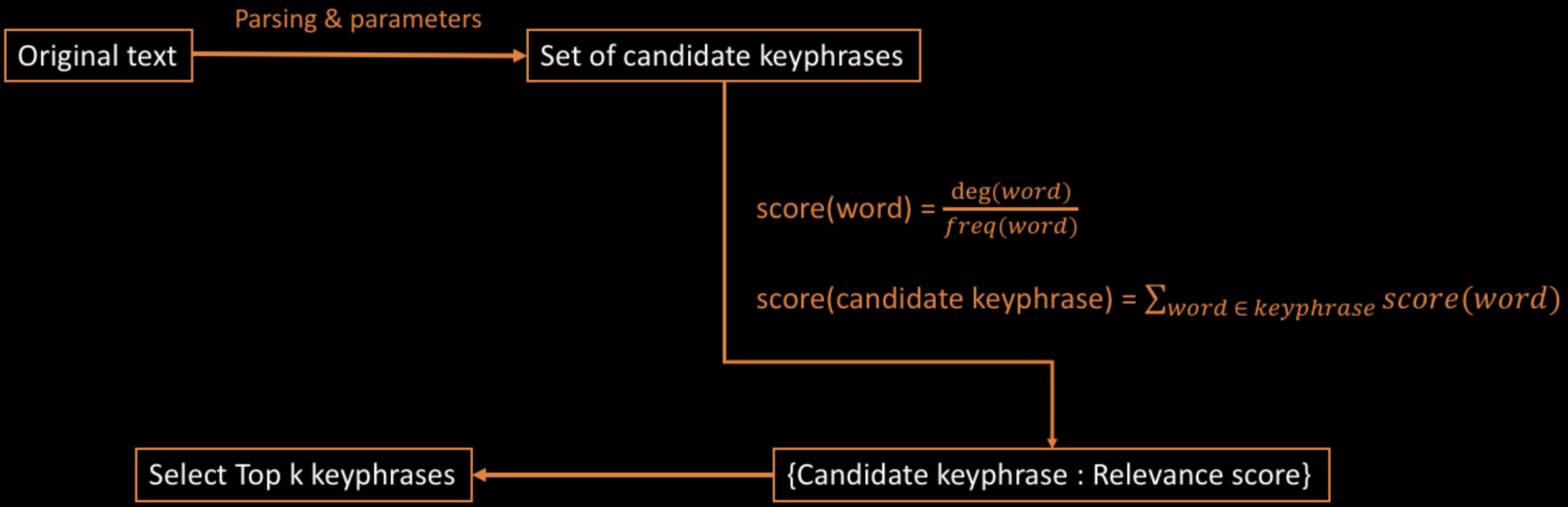
Output

- List of keyphrases and the associated relevance score

Keyphrase extraction

RAKE algorithm

Pipeline



Original text

Parsing & parameters

Set of candidate keyphrases

$$\text{score(word)} = \frac{\deg(word)}{freq(word)}$$

$$\text{score(candidate keyphrase)} = \sum_{word \in keyphrase} score(word)$$

{Candidate keyphrase : Relevance score}

Select Top k keyphrases

Keyphrase extraction

Modifications and improvements

Spell check
mosturizd -> moisturized

Use of stemming to group similar keyphrases
{moisturized, moisturizing} -> moistur

Negations aggregation
not oily -> no_oily / 2 words -> 1 word

Added parameter
minimum length of a keyphrase

Score Scaling

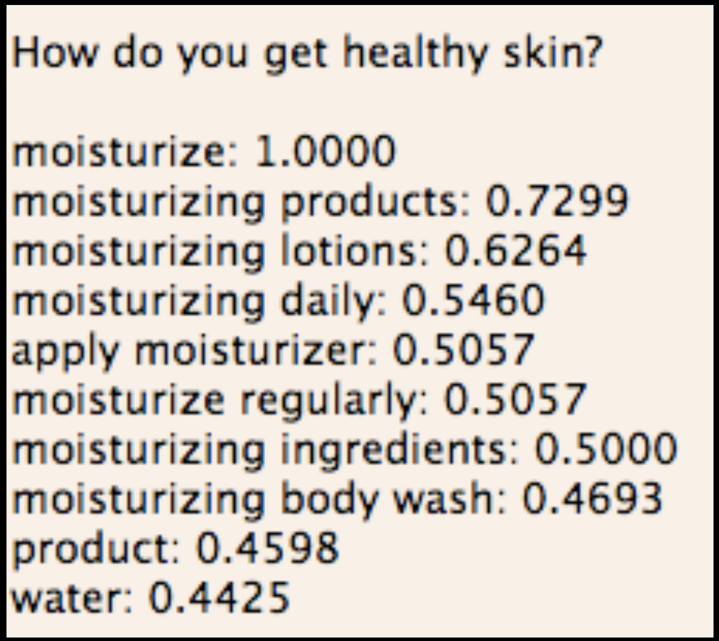
New scoring function

Keyphrase extraction

Modifications and improvements

New scoring function

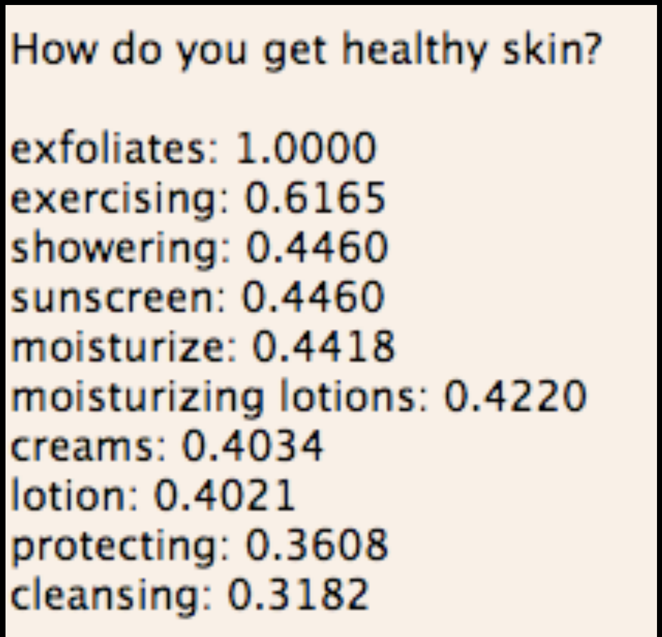
score(word) = $freq(word)$



```
How do you get healthy skin?

moisturize: 1.0000
moisturizing products: 0.7299
moisturizing lotions: 0.6264
moisturizing daily: 0.5460
apply moisturizer: 0.5057
moisturize regularly: 0.5057
moisturizing ingredients: 0.5000
moisturizing body wash: 0.4693
product: 0.4598
water: 0.4425
```

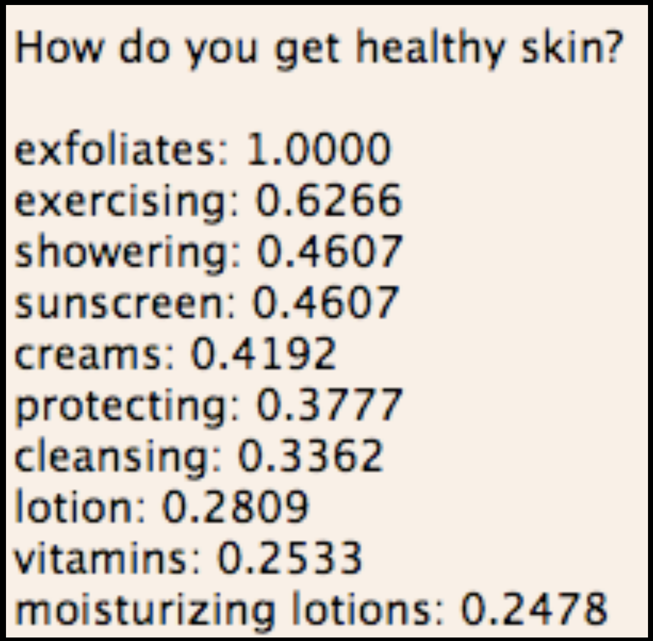
score(word) = $\frac{freq(word)}{\deg(word)^{0.8}}$



```
How do you get healthy skin?

exfoliates: 1.0000
exercising: 0.6165
showering: 0.4460
sunscreen: 0.4460
moisturize: 0.4418
moisturizing lotions: 0.4220
creams: 0.4034
lotion: 0.4021
protecting: 0.3608
cleansing: 0.3182
```

score(word) = $\frac{freq(word)}{\deg(word)}$

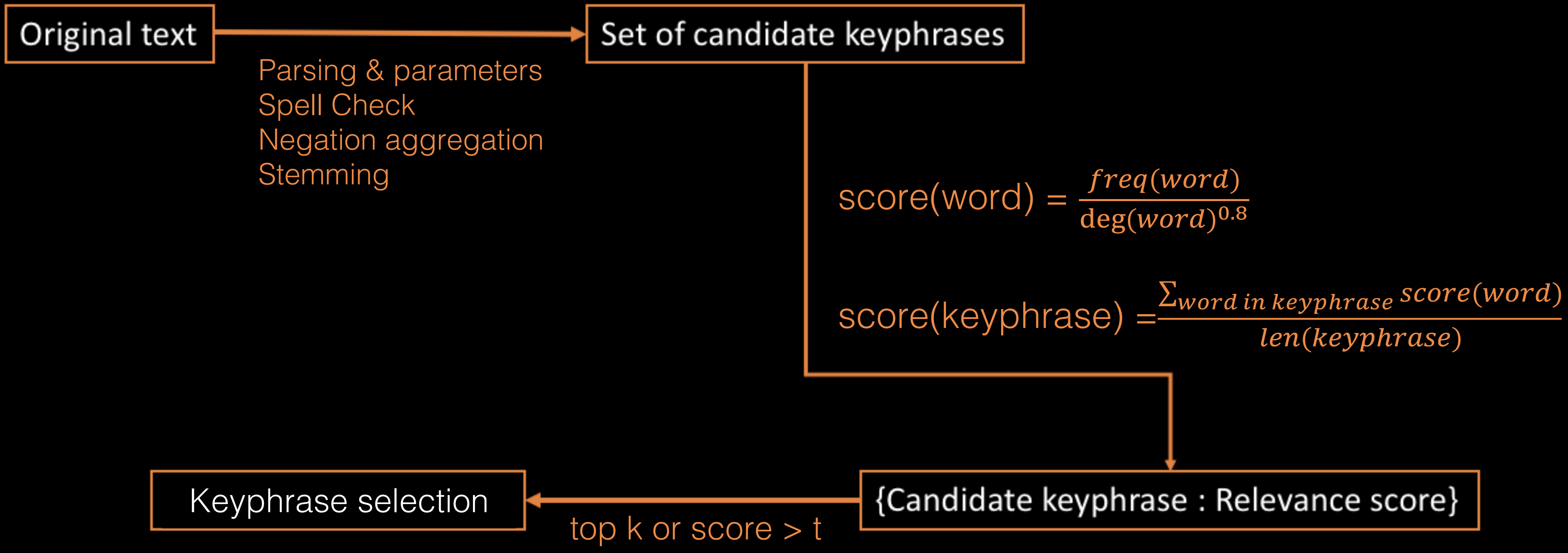


```
How do you get healthy skin?

exfoliates: 1.0000
exercising: 0.6266
showering: 0.4607
sunscreen: 0.4607
creams: 0.4192
protecting: 0.3777
cleansing: 0.3362
lotion: 0.2809
vitamins: 0.2533
moisturizing lotions: 0.2478
```

Keyphrase extraction

Final Pipeline



Original text

Parsing & parameters
Spell Check
Negation aggregation
Stemming

Set of candidate keyphrases

$$\text{score(word)} = \frac{freq(word)}{\deg(word)^{0.8}}$$

$$\text{score(keyphrase)} = \frac{\sum_{word\ in\ keyphrase} score(word)}{len(keyphrase)}$$

{Candidate keyphrase : Relevance score}

top k or score > t

Keyphrase selection

User Interface

Backend & algorithms

- Python
- Python Flask

Frontend

- HTML & CSS
- Javascript

Hosted on Heroku
Available at https://unilever-nlp.herokuapp.com/

Live Demo

Future developments

User-friendliness of the interface

Refine the algorithms

Other features

- Display the original sentence by clicking on a keyphrase
- More information per keyphrase: frequency, word scores…
- Export function
- Copy Paste input

Conclusion

Hard skills

Relationship with the mentors

Industry problem

Industry solution

Day at Unilever’s Office

Questions?

Contact Info

Ian Johnson: icj2103@columbia.edu

Julien Maudet: jm4418@columbia.edu